Hu, Rui; Yan, Zheng; Ding, Wenxiu; Yang, Laurence T.

A survey on data provenance in IoT

Published in:
World Wide Web

DOI:
10.1007/s11280-019-00746-1

Published: 01/03/2020

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Hu, R., Yan, Z., Ding, W., & Yang, L. T. (2020). A survey on data provenance in IoT. *World Wide Web*, *23*(2), 1441-1463. https://doi.org/10.1007/s11280-019-00746-1

World Wide Web
https://doi.org/10.1007/s11280-019-00746-1

A survey on data provenance in IoT

Rui Hu[1] · Zheng Yan[1,2] · Wenxiu Ding[1] · Laurence T. Yang[3]

Received: 23 January 2019 / Revised: 5 August 2019 / Accepted: 9 October 2019
Published online: 27 November 2019

Abstract
Internet of Things (IoT), as a typical representation of cyberization, enables the interconnection of physical things and the Internet, which provides intelligent and advanced services for industrial production and human lives. However, it also brings new challenges to IoT applications due to heterogeneity, complexity and dynamic nature of IoT. Especially, it is difficult to determine the sources of specified data, which is vulnerable to inserted attacks raised by different parties during data transmission and processing. In order to solve these issues, data provenance is introduced, which records data origins and the history of data generation and processing, thus possible to track the sources and reasons of any problems. Though some related researches have been proposed, the literature still lacks a comprehensive survey on data provenance in IoT. In this paper, we first propose a number of design requirements of data provenance in IoT by analyzing the features of IoT data and applications. Then, we provide a deep-insight review on existing schemes of IoT data provenance and employ the requirements to discuss their pros and cons. Finally, we summarize a number of open issues to direct future research.

Keywords cyberization · internet of things · distributed architecture · intelligent service · data provenance

This article belongs to the Topical Collection: Special Issue on Smart Computing and Cyber Technology for Cyberization
Guest Editors: Xiaokang Zhou, Flavia C. Delicato, Kevin Wang, and Runhe Huang

✉ Zheng Yan
zheng.yan@aalto.fi

[1] The State Key Laboratory of ISN, School of Cyber Engineering, Xidian University, Xi'an 710071 Shaanxi, China

[2] Department of Communications and Networking, Aalto University, 02150 Espoo, Finland

[3] Department of Computer Science, Francis Xavier University, Antigonish, NS B2G 2W5, Canada

1 Introduction

Cyberization refers to using communication and computer technologies to interconnect computers and various electronic terminal devices distributed in different locations. It allows users to share software, hardware and data resources according to certain network protocols. Cyberization has greatly improved the practical utility of computers and has been widely applied in transportation, finance, business management, education, telecommunications, commerce, and so on in our daily life. It stimulates the revolutionary development in both industry and scientific research and brings great convenience to human life. Internet of Things (IoT), as a typical representation of cyberization, connects all physical objects to the network by enabling the exchange of information between physical objects and a cyber world, which can also be applied in various domains by providing advanced and intelligent services. However, cyberization makes IoT encounter various security and privacy challenges [11, 17, 34, 58]. In an IoT system, it is difficult to perform identity management, guarantee the trustworthiness of data, detect abnormal behaviors and control the access to various data in IoT.

Data provenance provides the capability to solve the aforementioned issues in IoT by recording information about data origins, data operations and processing history from its source to current state. Thus, it becomes possible to track the sources or origins of any problems in IoT. Bauer and Chreckling [9] firstly proposed the application of data provenance in IoT, stated the requirements of applying data provenance in IoT as well as designed a conceptual model as a common architecture of data provenance in IoT. Since then, more and more researchers started to pay attention to data provenance in IoT and some provenance management schemes were proposed and applied in various intelligent IoT services. However, the literature still lacks a comprehensive survey on data provenance in IoT.

In this paper, we concentrate on data provenance in IoT and perform a comprehensive survey on existing related works for the purpose of pointing out open issues to instruct future research directions. Generally, the main contributions of our survey can be summarized as follows:

- We propose uniform criteria on data provenance in IoT by analyzing IoT distributed architecture and reviewing data provenance techniques and applications.
- We review existing works about data provenance in IoT and analyze their pros and cons according to our proposed criteria and security requirements.
- We point out a number of open issues based on the thorough survey and attempt to direct future research trends in the field of data provenance in IoT by evaluating its recent advance.

The rest of this paper is organized as follows. In Section 2, we introduce the architecture of IoT and the development of data provenance, followed by its application scenarios in Section 3. In Section 4, we point out the challenges to achieve data provenance in IoT and specify its uniform requirements. Then, we give a detailed overview and discussion on existing works in Section 5. Furthermore, we come up with several open issues and indicate future research directions in Section 6. Finally, a conclusion is drawn in the last section.

2 Background

In this section, we introduce the typical architecture of IoT, the basic concepts of data provenance, its main usages and implementation technologies.

2.1 Internet of things

IoT realizes the connection of any objects at any time and any place, which helps realize the organic integration of the human society and the physical world. It enables human beings to manage their appliances and improve their lifestyle in an elaborate and dynamic way. To implement IoT smart services, distributed IoT architecture has been proposed and applied, which is a general three-layer architecture as described in [59]. Al-Fuqaha *et.al* [2] extended the three-layer architecture to a five-layer architecture. An IoT architecture including a middle layer was presented in [29]. In what follows, we introduce the middle-layer into a general three-layer architecture and come up with a four-layer IoT architecture to understand data provenance in IoT, as shown in Fig. 1. In addition, we indicate security issues and existing solutions according to the requirements specified in [67].

Perception layer, also called physical layer, is the base layer of IoT, which includes numbers of devices embedding sensors and actuators to collect various information from a surrounding environment. The information can be time, temperature, humidity, voice, image, Quick Response (QR) code, and so on. Then, the devices send the collected information to a network layer. Since collected information in the perception layer may come from a variety of nodes, it is difficult to prevent malicious nodes from inserting invalid information. Moreover, the public environment threats data integrity and confidentiality. To solve these issues, some algorithms and authentication mechanisms have been applied in the perception layer to ensure data integrity and confidentiality [18].

Network layer, also called transmission layer, mainly forwards the collected data from physical objects to an information processing system. In this layer, data can be transferred through wired or wireless connections such as WiFi, Bluetooth, infrared, etc. Besides the attacks towards data integrity and confidentiality in a traditional network, data compatibility, privacy and cluster security problems may also occur in the network layer due to heterogeneity and complexity of IoT. Thus, secure routing protocols and data protection schemes need to be designed in this layer for protecting data security in IoT [1].

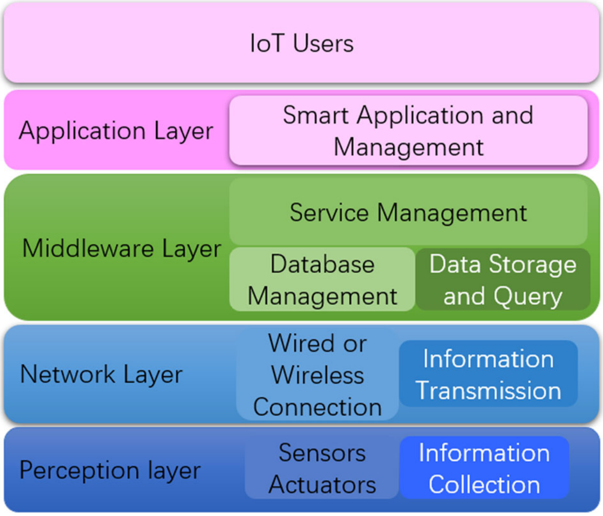

Figure 1 An IoT architecture

Middleware layer is responsible for service management and data processing. It also links to the database to store received data from the network layer by applying such typical techniques as virtualization and cloud computing.

Application layer is the interface between an IoT system and its users, which offers smart applications and corresponding management in order to provide intelligent services to IoT users. In this layer, different security requirements such as data access and privacy protection may raise in different applications. Generally, secure access control is one common concern in the application layer.

2.2 Data provenance

In fact, provenance is a kind of metadata that records the creation and usage of some entity, which is also called lineage, pedigree, genealogy [47]. Since data provenance was studied in a heterogeneous database system by Wang et al. in 1990s [54], some researchers have paid attention to data provenance in information systems and gave similar definitions [12, 22, 30, 56]. To sum up, data provenance was defined as the whole process information of data generation and evolution over time, including the static origins of data and their dynamic evolution [20].

Herein, we distinguish two concepts: *data provenance* and *provenance data/information* as below. *Data provenance* refers to a method or technique that can be used to record data origin and/or data operation and processing history in various applications. *Provenance data* is a kind of metadata that records data origin and/or data processing history that can be collected by a data provenance technique.

2.2.1 Usages of data provenance

Through collected provenance information, we can ascertain the original node that produced data, the immediate node that the data passed through, the user who is relevant to the data and the operations performed on the data. Sometimes we can even know the time and location of the behaviors acted on the data. Thus, data provenance is widely applied in scientific and business domains.

There are several main usages of data provenance. The data provenance can be used to assess the quality and trustworthiness of data. If the source and immediate nodes producing or transferring the data are trustworthy, the data quality and trustworthiness can be ensured [44]. In addition, data provenance can be used to detect errors in data generation and processing and find out the nodes and the actions that produced the errors, which can be used in verifiable computation [57, 61, 62]. On the other hand, detailed provenance information allows data recovery when data is unusable to maintain system availability and ensure smooth data communications [19]. Finally, data provenance can describe data citations and improve data readability [24].

2.2.2 Data provenance technologies

Currently, the technologies used for data provenance can be divided into three categories: logging-based, cryptography-based and blockchain-based.

1) Logging-based technologies

Logs are generated by pre-designed system programs during process creation and system calls, which can record data transfer and usage in a system. Traditional logs only record data in a single node, which limits the applications of logging to such scenarios as cloud computing and resource virtualization [37]. To overcome the limitation of traditional logs, data-centric end-to-end provenance mechanism (S2Logger) was designed for cloud computing in [49]. It can record the allocation of resources and data transmission between hosts and virtual machines. Though logging is an efficient provenance technique to monitor system events and end-to-end data transmission, it is inefficient to track data that pass through multiple nodes in a network and not applicable for provenance in distributed systems (e.g., IoT applications).

2) Cryptography-based techniques

Cryptographic mechanisms such as Message Authentication Code (MAC) and digital signatures on data can be applied to identify the origin of data, which can provide incomplete data provenance [21]. They can quickly verify data sources no matter how many nodes the data pass through. However, cryptographic methods cannot record the data processing history. In addition, data may come from various devices and nodes, which makes key management difficult.

3) Blockchain-based techniques

Blockchain is a decentralized ledger maintained by all peers in a peer-to-peer network, which can offer distributed data storage. Thus, it has been applied to provide data provenance by recording data operations in the form of blockchain transactions [45].

ProvChain is a blockchain-based provenance system that views blockchain as a distributed database to ensure data integrity and verifiability [32]. Once an operation is performed on data, the system stores a piece of record in a local database and uploads a provenance entry into a blockchain network. A Provenance Auditor (PA) can retrieve all provenance data from the blockchain network and maintain the local database. Figure 2 shows the interaction framework of ProvChain.

To achieve automatic blockchain-based provenance without provenance information uploading and off-chain verification, several scholars introduced smart contracts into the provenance system [39, 43]. Smart contracts include a function for tracking data changes and define access rules for data. With the specific functions defined in smart contacts, the privacy of shared data can be guaranteed.

Though aforementioned provenance approaches have proved the applicability of using blockchain for data provenance, limitations still exist due to the complexity of blockchain.

We summarize advantages and disadvantages of the above three types of data provenance technologies in Table 1. In addition to the three main technologies, provenance such as data routing information can be transferred along with main data by embedding them into a Bloom filter [14]. In addition, some unique features of devices (e.g., physical unclonable functions) can also be used to identify data source [50].

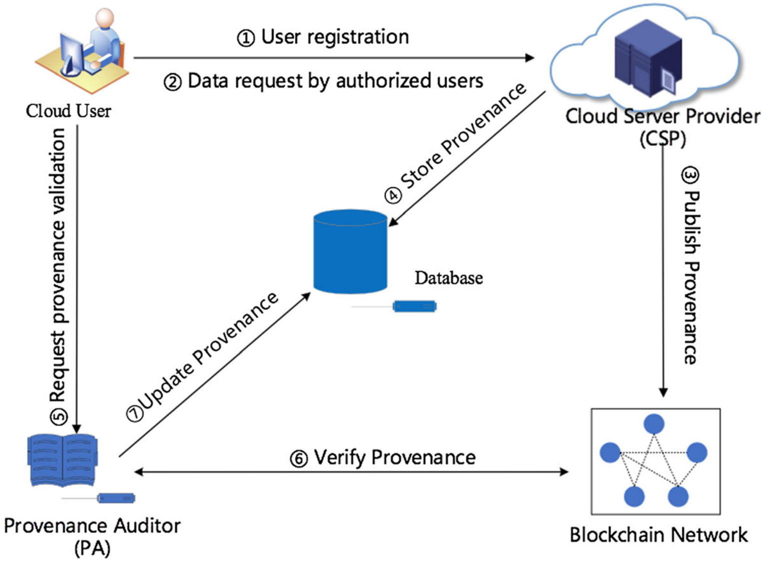

Figure 2 ProvChain interaction framework

3 Application scenarios of data provenance

Because data provenance can record data origin, evolution and transmission process in detail, it can be applied into multiple application scenarios to improve data service quality and support data auditing. According to the distributed IoT architecture and the data provenance applications in different layers of IoT, this section discusses some typical application scenarios of data provenance: distributed applications, Wireless Sensor Networks (WSN), cloud computing and IoT smart applications.

3.1 Distributed applications

Since the IoT can also be treated as a distributed application, researching the data provenance in distributed applications may provide some guidelines for studying data provenance in the IoT.

Distributed applications refer to assigning a task to different computers or servers through networks. In order to apply the data provenance to distributed systems or applications, some scholars proposed advanced and standardized solutions for provenance capture and collection [42, 48]. Raju et al. [42] designed a Producer API (PAPI) to work in distributed environments by means of unique object identification. The main advantages of PAPI are collecting

Table 1 Comparison of data provenance technologies

Technologies	Advantages	Disadvantages
Logging	Easy to operate and manage in a system	Limited in distributed applications
Cryptography	Able to identify the origin of data efficiently by certifying digital signature	Lack of data processing history records
Blockchain	Adaptable to distributed systems	Complex implementation

provenance in desired granularity and supporting production, capture and analysis of provenance information in a distributed application environment. It makes the provenance collection become more effective than before. In [48], the focus was to develop techniques to enrich simplified provenance information and achieve the completeness of provenance. Warekuromor et al. [55] advocated an extensible bottom-up distributed approach as the core of the spatial data infrastructure and demonstrated its feasibility. The advantage of this approach is that the spatial data infrastructure can grow organically without the involvement of a central entity, so the spatial data can be obtained quickly, which brings environmental and economic benefits.

3.2 WSN

The data collected by sensors in WSN may be used for decision-making in many domains such as supervisory control, battlefield monitoring and e-healthcare. Therefore, assessing the trustworthiness of collected data is quite crucial. Data provenance plays an important role in evaluating the trustworthiness of data. However, there are several challenging requirements of data provenance in WSN, which includes low energy and bandwidth consumption, secure transmission and efficient storage. Driven by these challenges, some secure provenance schemes in WSN have been proposed accordingly [27, 33, 52].

Lim et al. [33] proposed a systematic method for assessing the trustworthiness of collected data based on data provenance. This approach prefers to regard the trust scores as a main factor to ensure the trustworthiness of data items by using data provenance to compute trust scores. The trust scores of data items are computed based on their value similarity and provenance similarity. This model is a direct application of data provenance by quantifying the provenance information, which takes it for granted that data provenance information is trustworthy. It also evaluates the efficiency and effectiveness of the proposed approach in terms of the computation of trust scores on the basis of the prerequisites that the evaluated data sets conform to a normal distribution.

Salmin et al. [52] introduced a secure provenance scheme in WSN by embedding the provenance information into a Bloom filter (BF) that is transmitted along with data. This scheme overcomes the problem of resource constraints. It only requires a single channel for the transmission of both data and provenance compared with previous proposals that need separate channels. At the same time, the proposed scheme satisfies some security requirements such as confidentiality, integrity and freshness. Confidentiality means that an attacker is not likely to gain information about the sensor nodes included in the provenance by observing data packets. Integrity indicates that the path nodes cannot be arbitrarily added or deleted by an adversary, e.g., unauthorized users, and freshness refers that replay attacks can be avoided.

Interoperability is also an issue in sensor Web provenance. Liang et al. [27] proposed an approach to support interoperability. The model extended the W3C PROV model, and used Web Ontology Language to encode and capture provenance. This paper introduces several requirements for sensor Web provenance that include spatiality, temporality, usability, and so on. It is worth noting that the integrated provenance information stated in this paper contains observation results, timestamp, the location where the observation happens, the sensor that results in observation and the reason why the observation is made.

IoT can collect data about physical objects through WSN, so introducing data provenance into WSN becomes essential. According to the review of the above researches, we can summarize several requirements including trustworthiness, efficiency, confidentiality and integrity, with which a data provenance management system should satisfy. We will introduce these requirements in detail in the next section.

3.3 Cloud computing

Provenance is particularly crucial in the cloud, because data in the cloud can be shared widely and anonymously. Without provenance, data consumers have no means to verify its authenticity or identity [38]. The use of cloud helps solving the effective computation, storage and analytics of big data [31, 64, 66]. Therefore, data security, privacy and management in the cloud have been concerned. Privacy-preserving data classification models were proposed to support smart IoT [64, 66]. Access control as a necessary step for managing data has been studied based on Access Control List (ACL), which is time consuming and scales poorly. Thus, some scholars attempt to enhance the secure and effective access control by applying data provenance.

A Cloud Provenance Authority data storage model was proposed to perform secure file access in the cloud by Mirajkar *et.al.* [36]. According to this model, a series of access control rules have been made based on the provenance information. A user who wants to access a file must has registered at the provenance management system firstly, which is used to prove that he is a legal user to gain the file access permission. This model is a direct application of data provenance in secure cloud data access control, but it does not mention how to implement the provenance management system.

Adam et al. [8] introduced a scalable architecture for managing cloud provenance by detaching provenance from its associated data. A chain of certificates is provided when a user requests provenance information according to the protocol. Then cloud provenance authorities can cooperate with each other to track data, thus it satisfies confidentiality and scalability. A proof-of-concept implementation was provided to show its practicability.

The two schemes introduced above mainly focus on secure access control of data stored in the cloud. Different from them, Liang et al. [32] built an architecture to collect and verify cloud data provenance by embedding the provenance data into a blockchain network, which enhances the privacy and availability of data in a cloud environment. The method of implementing data provenance based on blockchain implies a future research direction of data provenance. From [32], we further notify that the requirements on privacy and availability of data provenance should be taken into consideration.

In addition to researches on the access control of cloud data storage, big data provenance analysis in cloud computing has also been concerned [15, 28]. Corresponding to IoT architecture mentioned in Section 2, the access control of secure provenance storage needs to be considered in the middleware layer. Hence, data provenance in the cloud with access control can be referred in the context of IoT data provenance.

3.4 IoT smart applications

In recent years, IoT applications are being developed rapidly and are becoming more and more popular in many domains owing to its intelligence with no need of any involvement of human beings through efficient and privacy-preserving big data analysis [31, 64–66]. IoT plays a critical role in many applications such as transportation and logistics, healthcare, smart home, and so on. On the other hand, the heterogeneity and mobility of IoT make its applications facing security and privacy issues [64, 66]. Thus, some researches applied data provenance into IoT to overcome these challenges. In this part, we list some literatures about applying data provenance in IoT smart applications.

Hossain et al. [23] focused on applying provenance to forensics investigations in an Internet of Vehicles environment, which is mainly used to ensure the integrity of collected and stored evidence. This work introduces an interaction provenance model, which records a chronologically ordered sequence of interactions exchanged by two or more actors. An investigator can ascertain the source and launching process of an attack through interaction provenance.

To meet with security, privacy and integrity requirements in an IoT healthcare monitoring system, Alshehri et al. [5] introduced a solution to establish trust management between the interacted sensors and devices by applying data provenance. Data provenance enables trustworthy communications among various devices and makes the exchanges of sensitive medical data between users and healthcare professionals secure. The authors proposed a data provenance module that states the origin, time stamp, location and historical manipulation of data.

To manage and aggregate a large volume of data collected from various devices in smart cities, some researchers proposed a new method to use provenance information to track device historical behaviors and data origin [10, 53]. Beran et al. [10] used a provenance model to design Trust Tiny Things system architecture, which enables transparent entity behaviors and avoids illegal leakage of personal data or other confidential data generated by connected devices. Different from Beran's proposal, reference [53] focuses on using sematic description and annotation to realize the fusion of heterogeneous data streams. In short, these two researches help users understand and trust collected data from various sources in smart cities.

In Table 2, we analyze and compare data provenance applications with regard to such aspects as the relationship with IoT, provenance information collected, services provided and techniques adopted to provide data provenance.

4 Data provenance in IoT

4.1 Architecture of data provenance in IoT

The interconnection of IoT brings great convenience to the society and human beings, but it also causes some security and privacy problems owing to data transmission over open networks. First, an attacker can easily insert fake data in the path of transmission and maliciously tamper data without being detected, which would reduce the trustworthiness and availability of data. Second, fake data spread fast, which can seriously affect multiple services. Third, data sharing and transmission, as well as processing increase the risk of user privacy leakage. Data provenance records the origin and processing history of the data in IoT, which provides the possibility to address the issues above.

Corresponding to the four-layer architecture of IoT as described in Section 2.1, a common architecture for data provenance in IoT is proposed and shown in Fig. 3. Provenance data as a kind of metadata can be collected in the physical layer and transmitted in the network layer, which is similar to main data. We consider adding the provenance management system to the middleware layer. Thus, the middleware layer should not only provide data storage and processing, but also store provenance information and support query on it.

Table 2 Application scenarios of data provenance

References	Application Scenario	Relationship with IoT	Provenance information	Service provided	Applied technology
[42, 48, 55]	Distributed applications	Same architecture as IoT	Data origin and processing history	Access control	Not mentioned
[27, 33, 52]	WSN	Physical layer and network layer of IoT	Data origin, time, location and route path	Assessing data quality	Bloom filter
[8, 32, 36]	Cloud computing	Middleware layer of IoT	Data source and processing history	Access control	Blockchain
[23]	IoT smart transportation	Application layer of IoT	Interaction provenance	Ensuring data integrity	Not mentioned
[5]	IoT smart healthcare	Application layer of IoT	Data origin, devices processing, time stamp and location information	Ensuring data trustworthiness	Not mentioned
[10, 53]	IoT smart city	Application layer of IoT	Data origin, time stamp, location information and processing history	Ensuring data trustworthiness	Not mentioned

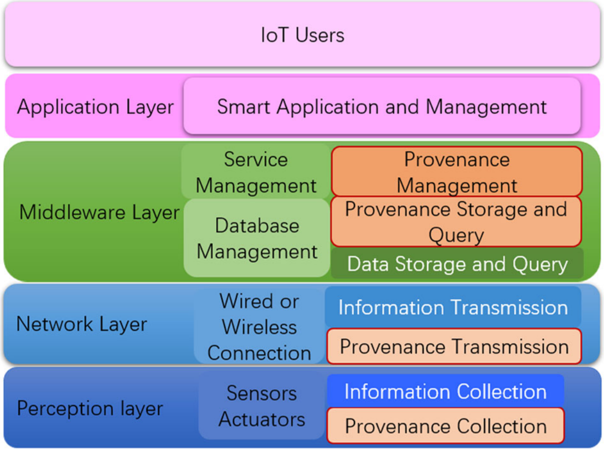

Figure 3 An IoT architecture to support data provenance

4.2 Challenges on data provenance in IoT

Though data provenance helps enhance the security of IoT data and improve the correctness and quality of IoT data analysis, it still faces a number of challenges when applying data provenance in IoT, which can be summarized as follows [4]:

1) The collection of provenance data. A quantity of data is collected and generated in IoT and may be processed multiple times in different services. In addition, many operations on different data may be performed at the same time in IoT. Hence, how to record all the sources and actions and generate sufficient provenance data without disclosing privacy becomes a significant issue.
2) The storage of provenance data. In an IoT system, original data may be transferred multiple times and have complex processing history, which may lead to a large size of provenance information. Thus, how to timely store and update the provenance data is also a challenging problem.
3) The flexible query of provenance data. When the faulty data is detected, the user wants to ascertain which node or which procedure causes the errors by inquiring provenance data. Due to the mobility of devices in IoT, it is expected that the query of provenance data should be highly flexible.
4) The security of provenance data. In an IoT system, provenance data may contain personal and sensitive information and may intrude original data security and user privacy. Hence, provenance should satisfy some security requirements, such as integrity, confidentiality, privacy protection, access control, freshness and availability [51]. Concretely, an adversary cannot dig out any information of original data by analyzing the provenance data and cannot modify any data sources and forwarding paths. In some special cases, the provenance data is more sensitive than original data, the privacy of provenance data should be guaranteed and only legal entities are allowed to access them. Ensuring the security of provenance data is vital.

In recent years, some data provenance schemes have been proposed or implemented regarding IoT. We will discuss them in detail in Section 5.

4.3 Requirements

In this subsection, we put forward some unified requirements on data provenance in IoT. We first list general requirements that each provenance system must meet. Then, we summarize essential security objectives that data provenance should satisfy in IoT.

4.3.1 General requirements

Completeness (CM): Colleting complete provenance information or data can fully take the advance to track data and actions for identity management, error detection, etc. As we all know, there are many actions and operations performed on data in IoT and all about data sources and operations should be recorded to acquire complete provenance information. Incomplete provenance information may lead to detection missing and suppression of some abnormal behaviors in IoT.

Trustworthiness (T): In order to apply provenance information to detect fake data and improve system security, we should first guarantee the trustworthiness of provenance itself. Otherwise, data provenance just complicates the system but fails to show its superiority.

Granularity (G): There are many types of data generated in IoT: data set, data packet, data flow and data file. For different data types, the information collected by the provenance management system varies in granularity. To be concrete, the provenance data or information may trace back to a component tuple of a data item, or record data source and processing history of related entities by viewing a data packet as a whole. In addition, the provenance of data flow may include interaction entities, interaction type, interaction time, and so on, which also regards the data flow as a whole. Moreover, not only the process derivation of a data file should be traced, the components of files such as paragraphs, shapes and images should be also traced with regard to their origins. In short, fine-grained provenance information helps achieving highly precise anomaly detection and auditing in IoT.

Depth (D): Data in IoT may be transmitted and processed for multiple times throughout its lifecycle. Moreover, the data is transmitted through multiple layers. Hence, it is worth considering to trace back to the original layer of data from its current state. For example, a data error was detected at the application layer in IoT, while the error may be generated at the physical layer or at the network layer. Thus, the depth of provenance about whether the source of the error can be traced back becomes one important requirement, which influences the accuracy of error detection.

Accuracy (A): Accurate positioning of abnormal behaviors can improve the efficiency of IoT systems and avoid wasting of resources. For example, in an IoT smart environment application, if a fire is suspected by fire sensors, the rescuers will locate the incident precisely through querying the provenance of alerts and arrive at the area of the incident as quickly as possible. On the contrary, the rescue efficiency becomes low if a provenance system provides a range of large area or informs an inaccurate location.

Efficiency (E): The collection and transmission of provenance information should not worsen the efficiency of the entire system. In addition, the query of provenance information should be as fast as possible and efficient provenance querying can avoid widespread error propagation.

Verifiability (V): Verifiability of provenance further enhances the reliability of provenance information. The end users can verify the provenance information to ensure that it is not modified. That is to say, any unauthorized participants or attackers cannot arbitrarily edit or update provenance information in the system.

Scalability (S): For managing the data collected and generated in IoT, the scalability of a provenance system need to be under consideration. Otherwise, with the increase of the data volume and the number of operations, it may become difficult to store complete provenance information, which possibly causes information loss.

4.3.2 Security requirements

On the premise of guaranteeing privacy and security properties in IoT, Aman et al. [6] proposed a novel and proper privacy-preserving data provenance mechanism in IoT. This paper points out that a privacy-preserving data provenance model must satisfy integrity and security. The provenance data should not disclose the identity of data owner to an illegal or unauthorized entity in terms of ensuring the confidentiality of provenance information. And the storage of provenance data is independent from data and provenance information can be verified. All collected provenance information especially sensitive data are not allowed to get out of the system. Thus, besides the general requirements as described in 4.3.1, security requirements must be satisfied, which are described as below.

Integrity (I): Once the provenance information is modified, identity management cannot be performed correctly, thus faulty data propagation is not possible to be blocked timely. To avoid an attacker or a set of cooperative attackers selectively adding or removing information into or from recorded provenance data, the integrity of provenance information should be ensured. In IoT applications, the integrity of provenance should be emphasized in the whole lifecycle of the data and the integrity includes source integrity and path integrity.

Confidentiality (CF): In the context of IoT, confidentiality implies that an adversary cannot gain any information about the provenance and data packet just by analyzing the packet data and metadata.

Freshness (F): At any layer of IoT, the captured provenance cannot be replayed. That is to say, the timeliness of provenance information should be guaranteed. Outdated provenance information cannot be used to figure out abnormal behaviors in IoT.

Privacy (P): Sometimes the metadata of data is more sensitive than the data. Especially in a dynamic and unprotected environment, the provenance, as a kind of metadata describing the data, also needs privacy protection. For example, in smart healthcare applications, the sources of healthcare data (i.e., the identity of a patient), need to be well protected.

5 Literature review and analysis

According to the criteria proposed in Section 4, we review the relevant works published in the past decade in this section, which are also summarized in Table 3. Herein, we focus on recent five-year works because few researchers paid attention to data provenance in IoT earlier.

Table 3 Comparison of existing methods based on general and security requirements

Reference	General Requirements								Security Requirements			
	CM	T	G	D	A	E	V	S	I	CF	F	P
[7]	*	–	Data packet	Y	–	–	Y	Y	Y	–	Y	–
[46]	*	Y	Data packet	Y	Y	N	Y	–	Y	–	–	–
[26]	–	–	–	–	–	Y	–	–	Y	–	–	–
[13]	Y	Y	–	–	Y	Y	Y	Y	–	–	–	–
[41]	Y	–	–	N	Y	Y	Y	Y	Y	–	Y	N
[25]	*	Y	Data packet	Y	Y	Y	Y	–	Y	Y	–	N
[3]	Y	Y	Data packet	Y	Y	Y	Y	Y	Y	Y	–	N
[35]	*	–	Data flow	Y	Y	Y	–	N	N	–	–	–
[40]	Y	–	Tuple	Y	–	Y	–	–	–	Y	–	–
[16]	*	Y	Data packet	Y	Y	Y	Y	N	Y	Y	Y	–
[60]	Y	–	–	–	Y	–	–	–	–	–	–	–
[63]	Y	–	Data flow	Y	Y	Y	–	Y	N	N	Y	N

Note:*: Partially satisfied; Y: Supported; N: Not supported; −: not mentioned

Through the introduction and analysis of data provenance techniques in Section 2.2.2, we can know that the current technologies applied for data provenance in IoT mainly include cryptography-based and blockchain-based methods. Thus, we focus on IoT data provenance applications based on these two types of methods. In addition, we also review a number of data provenance frameworks based on Physical Unclonable Functions (PUFs), programming achievement or some specific techniques.

5.1 Cryptography-based methods

As we analyzed in Section 2.2.2, cryptography-based data provenance in IoT can only record data origin, which provides incomplete provenance information. Thus, the following reviewed three schemes cannot support the completeness of provenance.

Baracaldo et al. presented a secure framework for protecting the sensitive data of provenance in IoT [7]. The proposed framework consists of four main modules: Policy Engine, Keyless Signature Infrastructure Module (KSI), IoT Platform Module and Auditing Service. The Policy Engine uses Attributed-Based Encryption (ABE) to ensure authorized users to gain access to the protected provenance data that are stored in an encrypted form, which provides the confidentiality and integrity of provenance. The KSI Module ensures the integrity of provenance. Policy Engine connects to a blockchain in order to protect and publish the top root of the KSI tree, which can achieve timeless update of provenance once the metadata has been updated. The IoT Platform Module serves as a manager for applications and the Auditing Service interacts with backend storage and builds connection between data points and provenance data. Because the collection of provenance data was not mentioned in this framework, we cannot judge the accuracy and trustworthiness of provenance.

In order to implement efficient and secure energy consumption management in smart home and avoid inaccurate messages or energy usage notifications being sent to home owners, a provenance scheme based on Shamir's secret sharing and RSA threshold cryptography was proposed in [46]. It uses signature to authenticate the source appliance of collected data. Which makes it possible to verify the provenance information. And only those who hold corresponding secret key can perform the verification, which achieves the confidentiality of provenance to a certain extent. RSA encryption algorithm guarantees the trustworthiness and integrity of provenance. In addition, the location of the data source can also be verified by Bluetooth beacons. Above all, the solution ensures the credibility of communications between sensors and applications. However, the complexity of cryptographic operations affects the efficiency of the proposed system.

Jayakody et al. [26] conducted a trust negotiation mechanism to ensure the integrity of shared provenance information. Before provenance data transmission, all nodes in IoT are configured with a key, which is shared between the peer nodes. When a piece of provenance message is transmitted from one node to another node, a random portion selected from the original provenance message was encrypted and transferred along with the original message. Upon receiving the message packet, the destination node extracts the cipher message portion and decrypts it with a pre-shared key in its possession. If the received provenance message portion matches the decrypted message, the integrity of the provenance is ensured. Selecting the partial message randomly encrypted by the shuffling algorithm makes the scheme efficient. Except for the integrity and efficiency of provenance, the solution does not consider other requirements.

5.2 Blockchain-based methods

Generally, blockchain-based provenance mechanisms for IoT can ensure the integrity and verifiability of provenance since every block in the blockchain network records the data operations in the form of transactions. And the creation and query of transactions determine the storage and query efficiency of provenance. In addition, the scalability of provenance depends on the storage capability of a blockchain network. In what follows, we review several related works.

A fully decentralized, blockchain-based traceability system AgriBlockIoT was designed by Caro et al. [13]. It creates transparent fault-tolerance, immutable and auditable records for Agri-Food supply chains. The proposed system can record the whole supply chain from product generation to product consumption. In other words, Blockchain is the main component of the system, which contains all business logics and provides consumers with the complete history of a purchased product. Every registered participant can store related operations in the blockchain and use the correct public/private key-pairs to digitally sign every operation. Thus, the system can collect complete provenance including data origins and operations performed on data, which support completeness of provenance. In addition, the use of smart-tags makes the record creation and retrieval effective.

In IoT, provenance information can also be used to meet trust concerns caused by the different devices in IoT edge cloud infrastructure [41], which records the origin of sensor data and other entities. An architectural pattern combing IoT edge orchestrations with a blockchain-based provenance mechanism was designed in [41]. All actions are recorded in the blockchain and provenance information can be stored in the blockchain network in the form of transactions. The consensus-driven mechanism of blockchain enables the verification of all the transactions in the network, i.e., the verification of provenance can be satisfied. Simultaneously, each transaction is hashed or digitally signed to maintain the integrity of provenance. Once a piece of data item is generated or an action is performed, a related transaction will be generated in the blockchain network. In addition, the W3C Provenance Model is used to represent provenance, which aids the assessment of quality, reliability or trustworthiness in data production and processing and ensures the accuracy of provenance. In this architecture, provenance information in the blockchain is open and it is uncertain whether an attacker can obtain any original data information through the analysis of provenance.

Based on blockchain variant smart contracts, Javaid et al. [25] designed a secure IoT framework called BlockPro, shown in Fig. 4. In BlockPro, each PUF produces a unique response for each device and uses it to identify the source of data. IoT devices first send their PUFs and corresponding responses to the blockchain and register through interaction with the server node in the blockchain network, which assures data provenance. Furthermore, a blockchain with two smart contracts (Smart Contract 1 and Smart Contract 2) is applied to ensure data integrity. Smart Contract 1 checks the establishment of data provenance to guarantee reliability of data origin. And Smart Contract 2 is responsible for storing and retrieving data provenance in the blockchain. The decentralized framework eliminates the need of any third parties and defends against a range of cyberattacks such as Denial of Service (DoS) and Distributed Denial of Service (DDoS). Through performance evaluation and simulation analysis, the authors proved the practicability and efficiency of BlockPro. However, BlockPro can only track data source without providing data processing history. Thus, it may not be applicable in some scenarios that request detailed provenance information.

To overcome the limitations of BlockPro, Ali et al. proposed a secure provenance framework for cloud-centric IoT, which not only identifies data origin but also provides periodic traffic profile of IoT devices [3]. The proposed framework consists of IoT devices, gateway nodes, cloud storage, and a blockchain network, shown in Fig. 5. IoT devices collect data and send them to a gateway node for processing. The cloud provides data storage, analysis and decision-making services. The blockchain was used to support data provenance in the form of transactions, which is different from BlockPro in terms of the design of the smart contracts. The blockchain ensures the integrity and confidentiality of provenance through smart contracts running in the blockchain. There are three types of transactions generated according to different smart contracts, which contain different information, initialized by different entities

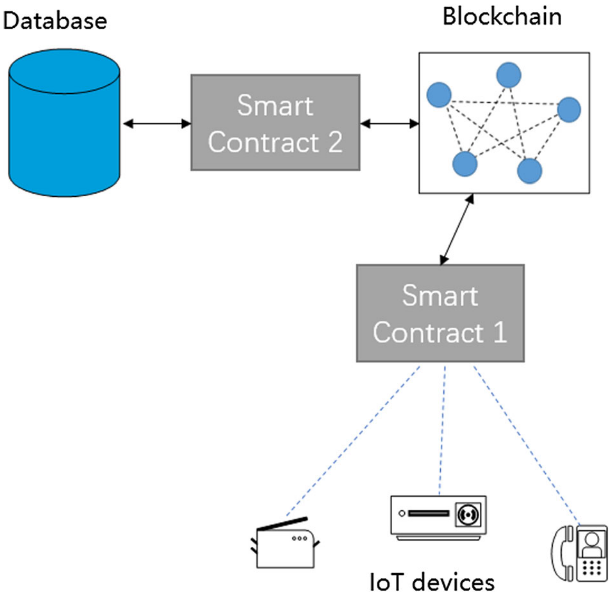

Figure 4 BlockPro framework

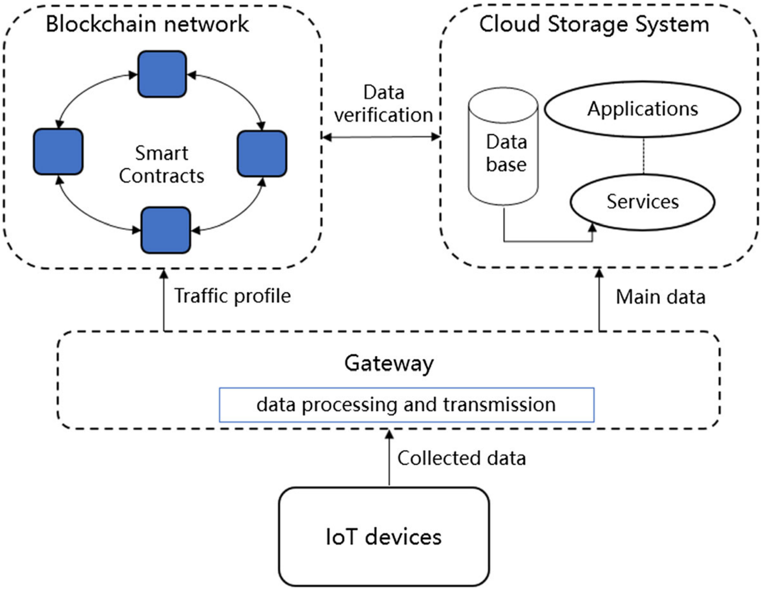

Figure 5 Secure provenance framework for cloud-centric IoT

and provided by different services. Moreover, the integrity and authenticity of IoT data with the digital signatures of a device can be verified by the gateway nodes. Furthermore, the blockchain can run in parallel with the cloud storage, which provides great efficiency.

5.3 Data provenance frameworks

Lomotey et al. [35] proposed a provenance scheme based on a lexical chaining technique to trace complicated data and detect faulty data propagation. A sample chain consists of a device, a type, an owner and communications. An IoT device is identified by a serial number of a particular type and is owned by a known user. In addition, the chain records the historical logs of previous communications between the device and other IoT devices. The adoption of Adjusted Rand Index (ARI) can help enhance the accuracy of estimation in the lexical chains. With the increase of the number of devices, the authors analyzed the synchronous and asynchronous communication means and response time in tablets and smartphones respectively to update detection cost. The effectiveness of the designed architecture was proved through experiments with scalability analysis. However, the security requirements were not considered in this scheme.

A provenance collection framework was introduced by Nwafor et al. [40]. In a provenance-sensor model, traced data can be converted into provenance information, and such conversion can happen at any layer of the IoT stack. A sensor trace can be defined as a tuple (t, e, a, s_1, r_1), where t represents a timestamp, e denotes an event (temperature, humidity), a indicates an operation (e.g., read, create, and update), t represents sensor information and r_1 denotes device information. With different configurations, the solution achieves data provenance for two different types of devices, which are the device with a single sensor and the device with multiple sensors, respectively. Thus, the provenance data collection system supports completeness. And an adversary cannot gain information by analyzing the trace tuple, so the provenance framework supports confidentiality. This system was implemented by using several tools and hardware components. For example, Raspberry Pi was used to demonstrate that the system is a low-cost and simple IoT demonstrator, and barectf is a light-weight generator of C code that generates trace data in Common Trace Format with satisfactory effectiveness. However, the integrity of the collected provenance information seems ignored in the system.

A preliminary data provenance protocol based on the physical layer in IoT was presented by Chia et al. [16]. The proposed protocol uses Physical Unclonable Functions (PUFs) and wireless link fingerprints to ensure self-trust and data provenance in IoT systems. The PUFs can identify data origin to achieve incomplete data provenance. And the wireless fingerprints ensure that the user can trust that the data has been collected from a stated location, which achieves the verification of data origin. A PUF is difficult to clone so the accuracy and integrity of the information origin can be ensured. A PUF is also sent from a specific device to a sever or gateway during the setup phase of the protocol. Thus, the origin of the data can be verified. The provenance data are sent along with the data, so its transmission is efficient and the provenance information cannot be replayed, satisfying the criterion of freshness. Based on a strong assumption that the PUFs are unique so that an adversary is difficult to copy. That is to say, an adversary cannot gain data information even though knowing the source of data, which supports confidentiality of provenance.

Yin et al. [60] designed a scheduling algorithm based data provenance for logistics chain in IoT. Provenance data about historical service information is recorded for each service provider. And the history of service reflects the trust and reputation of the server provider. In order to provide reliable and effective logistic transportation, the server provider that has high reputation can be chosen from a service provider pool to take part in the service chain. The provenance information is quantified in the designed algorithm, thus the accuracy of provenance information should be satisfied. But the completeness of provenance, the granularity of provenance, the depth of provenance, the efficiency of provenance, the verifiability and scalability of provenance were not discussed in this paper.

Due to the lack of context-aware services in IoT, Zhang et al. introduced workflow provenance to record the usage history and historical service behaviors of sensor devices [63]. A workflow in IoT possibly involves multiple sensors to provide services and is connected to a virtual device that comprises a finite set of actions for each sensor service of the workflow. Through establishment and analysis of workflow-sensor networks, historical service behavior information can be obtained and help the sensor service discovery to select proper sensors to provide considerate services. The authors decomposed the process of sensor service discovery into two phases: local optimization and global optimization that are illustrated by corresponding algorithms. In addition, a sensor application server was developed to achieve workflow provenance management and sensor device discovery and composition, which provides scalability and freshness in provenance based on Message-Bus. The trustworthiness of the workflow provenance depends on the construction of the virtual device. The accuracy and availability of the model are measured in performance evaluation based on this provisioning infrastructure. However, a malicious user may obtain historical service information and even modify sensor data by attacking virtual devices, which threatens the integrity and confidentiality of provenance. In addition, the integrity of sensor data provenance was not in consideration and no entities can verify the accuracy of historical behaviors.

6 Outlook of open issues and future directions

6.1 Open issues

Although applying the data provenance into IoT can help solving its security, trust and privacy issues, there are still few researches performed in this research field. Most of the proposed schemes focus on the concrete applications of data provenance. Few proposals consider how to implement a general data provenance management system. And there is almost no system that fully implements the data provenance in IoT by satisfying all proposed general requirements and security requirements. The existing schemes still suffer from the following problems.

First, processing and storage of huge amount of data were not considered in most of existing work. Most of existing schemes are based on the assumption that base stations have unlimited capabilities or there is only a single server considered. Thus, the issues to process and store big data has not been well investigated and solved.

Second, attachment of provenance data to data item impacts the flexibility of data provenance management. In the reviewed schemes, some use digital signatures to achieve data source authentication. Some attach transformation history and related node information to the data and send the provenance information along with the data to immediate nodes and destination nodes. To some extent, these methods are efficient and save resource consumption. But, many existing methods of provenance data attachment reduce the flexibility of data provenance management. A more flexible method is highly expected.

Third, privacy of provenance has not yet been well solved. As an important basis, privacy protection of provenance information has not been well considered in most of IoT data provenance management systems. In fact, privacy is hard to be fully protected by only providing efficient and secure access control.

Fourth, holistic provenance management that satisfies all proposed requirements is still missed. As presented in Table 3, no scheme satisfies all the general requirements and security requirements. Almost no scheme takes into account the distributed architecture of IoT, decentralization and scalability are not well supported.

6.2 Future research directions

Based on the discovered open issues, we move up to point out four significant directions to instruct future research.

First, studying data fusion and filtering mechanisms can greatly help solving the issues caused by big provenance data. To handle the collection and storage of large-scale provenance data, it is significant to use data fusion and classification techniques to compress the collected provenance data into small sized data according to some principles or rules, such as original devices and all records of one product. In addition, data filtering according to different requirements should be explored to extract useful and expected provenance data. Herein, we need to study reversible techniques to get original data in case tracking is needed, even though the data is fused or highly compressed.

Second, flexible data provenance management is expected to be realized in order to support different and various real applications. Two possible ways to solve this problem: 1) for the data with less operations performed, the attachment of provenance information to main data can be considered; 2) the provenance information should be managed separately to avoid impacting the transmission efficiency of main data if a large amount of operations will be performed on the main data. But an ideal solution could be a set of provenance data can be flexibly used for supporting multiple IoT applications at the same time based on real demands. How to achieve this with high efficiency and economy is an interesting research topic.

Third, the privacy protection of provenance data should be considered. Traditional privacy protection methods and technologies may not be applicable to preserve provenance data, new methods should be investigated to support privacy-preserving data provenance in IoT.

Fourth, research of a distributed data provenance architecture becomes essential in IoT owning to its specific characteristics and requirements. To achieve a secure and comprehensive data provenance management system in IoT, it is interesting to apply some decentralized technology, such as blockchain, and integrate it with cryptographic methods to meet the requirements of data provenance in IoT in order to build up a distributed architecture for IoT data provenance. Herein, how to protect the privacy of provenance information becomes an interesting topic to explore.

7 Conclusion

In this paper, to tackled the security and privacy issues in cyberization process, we introduced the basic of data provenance in IoT and figured out a number of general requirements and security requirements by overviewing IoT data provenance applications. According to the proposed requirements, we provided a deep-insight review on existing methods of IoT data provenance. Based on the review and analysis, we summarize some open issues to direct future research. Although data provenance plays an important role in IoT, there still exist challenges, such as big data handling, privacy preservation, flexibility support and decentralization, which motivates future efforts in this research field.

Acknowledgements Open access funding provided by Aalto University. This work is supported in part by the National Natural Science Foundation of China under Grants 61672410, 61802293 and U1536202, the National Postdoctoral Program for Innovative Talents under grant BX20180238, the Project funded by China Postdoctoral Science Foundation under grant 2018 M633461, the Academy of Finland under Grants 308087 and 314203, the Key Lab of Information Network Security, Ministry of Public Security under grant No. C18614, the open grant of the Tactical Data Link Lab of the 20th Research Institute of China Electronics Technology Group Corporation, P.R. China under grant CLDL-20182119, Shaanxi Innovation Team Project under grant 2018TD-007, and the 111 project under grant B16037.

References

1. Airehrour, D., Gutierrez, J., Ray, S.K.: Secure routing for internet of things: a survey. J. Netw. Comput. Appl. **66**, 198–213 (2016)
2. Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., Ayyash, M.: Internet of things: a survey on enabling technologies, protocols, and applications. IEEE Communications Surveys & Tutorials. **17**, 2347–2376 (2015)
3. S. Ali, G. Wang, M. Z. A. Bhuiyan, and H. Jiang, "Secure Data Provenance in Cloud-Centric Internet of Things Via Blockchain Smart Contracts," in IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, 2018, pp. 991–998

4. Alkhalil, A., Ramadan, R.A.: IoT data provenance implementation challenges. Procedia Computer Science. **109**, 1134–1139 (2017)
5. M. Alshehri, M. Elkhodr, and B. Alsinglawi, "Data Provenance in the Internet of Things," in Proceedings of 32nd International Conference on Advanced Information Networking and Applications Workshops, 2018, pp. 727–731
6. M. N. Aman, K. C. Chua, and B. Sikdar, "Secure data provenance for the internet of things," in Proceedings of the 3rd ACM International Workshop on IoT Privacy, Trust, and Security, 2017, pp. 11–14
7. N. Baracaldo, L. A. D. Bathen, R. O. Ozugha, R. Engel, S. Tata, and H. Ludwig, "Securing Data Provenance in Internet of Things (IoT) Systems," in International Conference on Service-Oriented Computing, 2016, pp. 92–98
8. A. Bates, B. Mood, M. Valafar, and K. Butler, "Towards secure provenance-based access control in cloud environments," in Proceedings of the third ACM conference on Data and application security and privacy, 2013, pp. 277–284
9. S. Bauer and D. Schreckling, "Data provenance in the internet of things," in Proceedings of 32nd International Conference on Advanced Information Networking and Applications Workshops, 2018, pp.727–731
10. S. Beran, E. Pignotti, and P. Edwards, "Trusted tiny things: Making devices in smart cities more transparent, " in Proceedings of the Fifth Workshop on Semantics for Smarter Cities-A Workshop at the 13th International Semantic Web Conference, 2014, pp. 83–95
11. E. Bertino, "Data Security and Privacy in the IoT," in Proceedings of the 19th International Conference on Extending Database Technology, 2016, pp. 1–3
12. P. Buneman, S. Khanna, and T. Wang-Chiew, "Why and where: A characterization of data provenance," in Proceedings of International conference on database theory, 2001, pp. 316–330
13. M. P. Caro, M. S. Ali, M. Vecchio, and R. Giaffreda, "Blockchain-based traceability in Agri-Food supply chain management: A practical implementation," in IoT Vertical and Topical Summit on Agriculture-Tuscany, 2018, pp. 1–4
14. A. M. Chacko, A. Cuzzocrea, and S. M. Kumar, "Automatic Big Data Provenance Capture at Middleware Level in Advanced Big Data Frameworks," in Connected Environments for the Internet of Things, 2017, pp. 219–239
15. A. Chacko and S. M. Kumar, "Big data provenance research directions," in Proceedings of Region 10 Conference, 2017, pp. 651–656
16. M. H. Chia, S. L. Keoh, and Z. Tang, "Secure Data Provenance in Home Energy Monitoring Networks," in Proceedings of the 3rd Annual Industrial Control System Security Workshop, 2017, pp. 7–14
17. Ding, W., Jing, X., Yan, Z., Yang, L.T.: A survey on data fusion in internet of things: towards secure and privacy-preserving fusion. Information Fusion. **51**, 129–144 (2019)
18. Fan, K., Gong, Y., Du, Z., Li, H., Yang, Y.: RFID secure application revocation for IoT in 5G. IEEE Trustcom/BigDataSE/ISPA. **1**, 175–181 (2015)
19. I. T. Foster, J.-S. Vöckler, M. Wilde, and Y. Zhao, "the Virtual Data Grid: a New Model and Architecture for Data-Intensive Collaboration," in CIDR, 2003
20. Gao, M., Jin, C.-Q., Wang, X.-L., Tian, X.-X., Zhou, A.-Y.: A survey on management of data provenance. Chinese Journal of Computers. **33**, 373–389 (2010)
21. G. Gibb, H. Zeng, and N. McKeown, "Outsourcing network functionality," in Proceedings of the first workshop on Hot topics in software defined networks, 2012, pp. 73–78
22. B. Glavic and K. R. Dittrich, "Data Provenance: A Categorization of Existing Approaches," in Datenbanksysteme in Business, Technologies und Web, 2007, pp. 227–241
23. M. Hossain, R. Hasan, and S. Zawoad, "Trust-IoV: A Trustworthy Forensic Investigation Framework for the Internet of Vehicles (IoV)," in IEEE International Congress on Internet of Things, 2017, pp. 25–32
24. Jagadish, H., Olken, F.: Database management for life sciences research. ACM SIGMOD Rec. **33**, 15–20 (2004)
25. U. Javaid, M. N. Aman, and B. Sikdar, "BlockPro: Blockchain based Data Provenance and Integrity for Secure IoT Environments," in Proceedings of the 1st Workshop on Blockchain-enabled Networked Sensor Systems, 2018, pp. 13–18
26. J. A. Jayakody, L. Rupasinghe, N. Mapa, T. Disanayaka, D. Kandawala, and K. Dinusha, "A Light Weight Provenance Aware Trust Negotiation Algorithm for Smart Objects in IoT." 2018
27. L. Jiang, W. Kuhn, and P. Yue, "An interoperable approach for Sensor Web provenance," in Proceedings of 6th International Conference on Agro-Geoinformatics, 2017, pp. 1–6
28. B. Jose, T. R. Ramanan, and S. M. Kumar, "Big data provenance and analytics in telecom contact centers," in Proceedings of Region 10 Conference, 2017, pp. 1573–1578
29. R. Khan, S. U. Khan, R. Zaheer, and S. Khan, "Future internet: the internet of things architecture, possible applications and key challenges," in Proceedings of 2012 10th International Conference on Frontiers of Information Technology, 2012, pp. 257–260

30. Lanter, D.P.: Design of a lineage-based meta-data base for GIS. Cartography and Geographic Information Systems. **18**, 255–261 (1991)
31. Li, P., Chen, Z., Yang, L.T., Zhang, Q., Deen, M.J.: Deep convolutional computation model for feature learning on big data in internet of things. IEEE Transactions on Industrial Informatics. **14**(2), 790–798 (2017)
32. X. Liang, S. Shetty, D. Tosh, C. Kamhoua, K. Kwiat, and L. Njilla, "Provchain: A blockchain-based data provenance architecture in cloud environment with enhanced privacy and availability," in Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, 2017, pp. 468–477
33. H.-S. Lim, Y.-S. Moon, and E. Bertino, "Provenance-based trustworthiness assessment in sensor networks," in Proceedings of the Seventh International Workshop on Data Management for Sensor Networks, 2010, pp. 2–7
34. D. Liu, Z. Yan, W. Ding, and M. J. I. I. o. T. J. Atiquzzaman, "A Survey on Secure Data Analytics in Edge Computing," IEEE Internet of Things Journal, 2019. **6**(3), pp. 4946–4967. https://doi.org/10.1109/JIOT.2019.2897619
35. Lomotey, R.K., Pry, J.C., Chai, C.: Traceability and visual analytics for the internet-of-things (IoT) architecture. World Wide Web. **21**, 7–32 (2018)
36. S. B. Mirajkar and S. Khatawkar, "A provenance-based access control model for securely storing data in cloud," in Proceedings of 2nd International Conference for Convergence in Technology, 2017, pp. 906–909
37. K. K. Muniswamy-Reddy, D. A. Holland, U. Braun, and M. Seltzer, "Provenance-aware storage systems," in Proceedings of Conference on Usenix 06 Technical Conference, 2006, pp. 43–56
38. K.-K. Muniswamy-Reddy, P. Macko, and M. I. Seltzer, "Provenance for the Cloud," in FAST, 2010, pp. 197–210
39. R. Neisse, G. Steri, and I. Nai-Fovino, "A blockchain-based approach for data accountability and provenance tracking," in Proceedings of the 12th International Conference on Availability, Reliability and Security, 2017, pp. 1–14
40. E. Nwafor, A. Campbell, D. Hill, and G. Bloom, "Towards a Provenance Collection Framework for Internet of Things Devices," in IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, 2017, pp. 1–6
41. C. Pahl, N. El Ioini, S. Helmer, and B. Lee, "An architecture pattern for trusted orchestration in IoT edge clouds," in Proceedings of 2018 Third International Conference on, 2018Fog and Mobile Edge Computing , 2018, pp. 63–70
42. B. Raju, T. Elsethagen, E. Stephan, and K. K. Van Dam, "A Scientific Data Provenance API for Distributed Applications," in Proceedings of International Conference on Collaboration Technologies and Systems, 2016, pp. 104–111
43. A. Ramachandran and D. Kantarcioglu, "Using Blockchain and Smart Contracts for Secure Data Provenance Management," CoRR, 2017
44. H. Sabaa and B. Panda, "Data authentication and provenance management," in Proceedings of 2nd International Conference on Digital Information Management, 2007, pp. 309–314
45. Salman, T., Zolanvari, M., Erbad, A., Jain, R., Samaka, M.: Security services using blockchains: a state of the art survey. IEEE Communications Surveys & Tutorials. **21**, 858–880 (2018)
46. J. L. C. Sanchez, J. B. Bernabe, and A. F. Skarmeta, "Towards privacy preserving data provenance for the Internet of Things," in Proceedings of IEEE 4th World Forum on Internet of Things, 2018, pp. 41–46
47. Simmhan, Y.L., Plale, B., Gannon, D.: A survey of data provenance in e-science. ACM SIGMOD Rec. **34**, 31–36 (2005)
48. E. Stephan, B. Raju, T. Elsethagen, L. Pouchard, and C. Gamboa, "A scientific data provenance harvester for distributed applications," in Scientific Data Summit, 2017, pp. 1–9
49. C. H. Suen, R. K. L. Ko, S. T. Yu, P. Jagadpramana, and S. L. Bu, "S2Logger: End-to-End Data Tracking Mechanism for Cloud Data Provenance," in IEEE International Conference on Trust, 2013, pp. 594–602
50. G. E. Suh and S. Devadas, "Physical unclonable functions for device authentication and secret key generation," in Proceedings of 2007 44th ACM/IEEE Design Automation Conference, 2007, pp. 9–14
51. S. Suhail, C. S. Hong, Z. U. Ahmad, F. Zafar, and A. Khan, "Introducing Secure Provenance in IoT: Requirements and Challenges," in Proceedings of International Workshop on Secure Internet of Things, 2016, pp. 39–46
52. Sultana, S., Ghinita, G., Bertino, E., Shehab, M.: A lightweight secure scheme for detecting provenance forgery and packet dropattacks in wireless sensor networks. IEEE Transactions on Dependable and Secure Computing. **12**, 256–269 (2015)
53. R. Tönjes, P. Barnaghi, M. Ali, A. Mileo, M. Hauswirth, F. Ganz, et al., "Real Time Iot Stream Processing and Large-Scale Data Analytics for Smart City Applications," in Proceedings of European Conference on Networks and Communications, 2014

54. Y. R. Wang and S. E. Madnick, "A polygen model for heterogeneous database systems: The source tagging perspective," in Proceddings of 16th International Conference on Very Large Data Bases, 1990, pp. 519–538
55. T. Warekuromor, A. James, B. Anifowose, and N. Trodd, "A distributed, scalable and provenance-enabled data access protocol for spatial data infrastructure," in Proceedings of IEEE 21st International Conference on Computer Supported Cooperative Work in Design, 2017, pp. 180–185
56. A. Woodruff and M. Stonebraker, "Supporting fine-grained data lineage in a database visualization environment," in Proceedings of 13th International Conference on Data Engineering, 1997, pp. 91–102
57. Yan, Z., Yu, X., Ding, W.: Context-aware verifiable cloud computing. IEEE Access. **5**, 2211–2227 (2017)
58. Yan, Z., Zhang, P., Vasilakos, A.V.: A survey on trust management for internet of things. J. Netw. Comput. Appl. **42**, 120–134 (2014)
59. Z. Yang, Y. Yue, Y. Yang, Y. Peng, X. Wang, and W. Liu, "Study and application on the architecture and key technologies for IOT," in Proceedings of 2011 International Conference on Multimedia Technology, 2011, pp. 747–751
60. J. Yin, J. Li, and P. Ke, "A Provenance Based Scheduling Algorithm for Logistics Chain in IOT," in Proceedings of 6th International Conference on Information Management, Innovation Management and Industrial Engineering, 2013, pp. 324–327
61. Yu, X., Yan, Z., Vasilakos, A.V.: A survey of verifiable computation. Mobile Networks and Applications. **22**, 438–453 (2017)
62. X. Yu, Z. Yan, and R. Zhang, "Verifiable outsourced computation over encrypted data," vol. 479, pp. 372–385, 2019
63. J. Zhang, N. Radia, Z. Li, N. Xin, Y. Ren, P. Sachdeva, et al., "An Infrastructure Supporting Considerate Sensor Service Provisioning," in Proceedings of IEEE 6th International Conference on Service-Oriented Computing and Applications, 2013, pp. 69–76
64. Q. Zhang, L.T. Yang, Z. Chen, P. Li, "PPHOPCM: Privacy-Preserving High-Order Possibilistic C-Means Algorithm for Big Data Clustering with Cloud Computing", IEEE Transactions on Big Data, 2017.: https://doi.org/10.1109/TBDATA.2017.2701816
65. Zhang, Q., Yang, L.T., Chen, Z., Li, P.: High-order Possibilistic C-means algorithms based on tensor decompositions for big data in IoT. Information Fusion. **39**, 72–80 (2018)
66. Zhang, Q., Yang, L.T., Chen, Z., Li, P., Deen, M.J.: Privacy-preserving double-projection deep computation model with crowdsourcing on cloud for big data feature learning. IEEE Internet Things J. **5**(4), 2896–2903 (2017)
67. K. Zhao and L. Ge, "A survey on the internet of things security," in Proceedings of Ninth international conference on computational intelligence and security, 2013, pp. 663–667

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.